Bioestadística para no estadísticos
Tome decisiones con criterio

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Capítulo 3:

Variabilidad:

El proceso de cuantificar la observación

Erik Cobo, José Antonio González y Pilar Muñoz
Jordi Cortés, Rosario Peláez , Marta Vilaró y Nerea Bielsa

Septiembre 2014

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

equator network

UPC

ELSEVIER DOYMA

MEDICINA CLINICA

TRIALS TRIALS

Variabilidad

Presentación

Sin variabilidad no hay vida. Y como "visto un caso, vistos todos", tampoco información. Dicho en positivo: sabemos distinguirnos porque sabemos procesar la variabilidad.

La Estadística aborda cómo recoger la información ("proceso de medida") y cómo representarla con gráficos y con números.

La primera parte de este tema razona sobre el significado de un valor observado; por ejemplo, ¿qué información aporta la etiqueta 'enfermo'? O, si vemos que alguien mide 178 cm: ¿es alto? ¿Es bajo? ¿Comparado con qué?

La segunda parte introduce los gráficos y los estadísticos más adecuados para cada tipo de variable, así como las medidas de posición relativa de las unidades. Todo ello se aplica a la descripción de los participantes en un estudio.

Contribuciones: (1) la versión original de marzo 2013 descansa en el libro de Bioestadística para No estadísticos de Elsevier de EC, JAG y PM, editada por JC y revisada por MV y R; (2) la de julio de 2013 fue revisada por MV, JC y EC para incorporar mejoras y sugerencias anónimas; y (3) la de septiembre de 2104 por NB y EC.

1. Medida

Una primera definición puede ser “medir es asignar números a objetos siguiendo reglas”.

Ejemplo 1.1: Asignación de 180 centímetros a Erik Cobo.

También podríamos asignar etiquetas —sin significado de número.

Ejemplo 1.2: Asignación de masculino a Erik Cobo.

Historieta: Malditas etiquetas que nos encasillan. Dice el Dr. Vives que él no es médico, que él es un corredor de fondo que trabaja como médico. Y yo digo que él es un gran médico.

Quedémonos con la idea de que esta asignación necesita reglas que la hagan reproducible. El proceso científico requiere establecer un lenguaje común, con idéntico significado para cualquier observador.

Historieta: En el lenguaje popular una misma frase puede tener diferentes perspectivas, matices o significados. Un popular “doble sentido” es saludar al amigo preocupado por su imaginaria calvicie con un: “¡Cuánto tiempo sin verte el pelo!”

Lecturas: El inicio de este tema sigue la línea de Stevens. Para una definición más formal, consulte la versión inglesa de Wikipedia.

1.1. Escala nominal

Clasificar consiste en agrupar los objetos estudiados: aquellos de la misma categoría deben ser equivalentes entre sí y diferentes de los de otra categoría.

Ejemplo 1.3: Una burda y primera clasificación de las personas las dividiría en enfermos y sanos. Dos enfermos compartirán ciertas características comunes que los hacen diferentes de los sanos.

Definición

La **escala nominal** clasifica a las unidades en grupos o categorías.

Nota: si todas las unidades fueran iguales, entonces todas pertenecerían a la misma categoría y no tendría sentido ni clasificarlas, ni medirlas.

Historieta: Un buen profesor consigue que todos sus estudiantes sean excelentes. Un mal evaluador pone a todos los estudiantes la misma nota.

Ejemplo 1.4: Los códigos de identificación personal pertenecen a la escala nominal. Incluso los formados por cifras, porque no tienen significado de número, ya que un valor 'mayor' no implica nada. El DNI solo "clasifica" y por tanto está en escala nominal –aunque especial: cada categoría sólo tiene 1 caso.

Ejercicio 1.1

Proponga ejemplos de otras clasificaciones posibles.

1.2. Escala ordinal

En el ejemplo anterior de enfermo y sano, la inmediata ambición del clínico y del científico es matizar más y, por ejemplo, establecer grados de intensidad: sano, leve, moderado y grave. Igual que antes, dos unidades de la misma categoría serán iguales entre sí y diferentes de las restantes categorías. Pero ahora, además, puede establecer una relación de orden y decir que grave es más que moderado; y como moderado es más que leve; entonces grave también es más que leve.

Definición

La **escala ordinal** cumple las propiedades de la escala nominal y, además, permite ordenar las categorías.

Ejemplo 1.5: El indicador BK de la tuberculosis puede valorarse en una escala ordinal que va desde 0 a 3 cruces (0 / + / ++ / +++).

Ejercicio 1.2

Proponga algún otro ejemplo de variable en escala ordinal.

1.3. Escala de intervalo

Una vez establecido un orden, la siguiente ambición del científico es comparar las diferencias entre categorías sucesivas. En el ejemplo de la tuberculina, ¿existe el mismo 'salto' de + a ++, que de ++ a +++? Si todos los 'saltos' tuvieran el mismo significado, se podría hablar de una misma **unidad de medida**, lo que permitiría comparar diferentes intervalos y decir, por ejemplo, que la diferencia entre + y +++ es mayor (el doble, como veremos) que la diferencia entre 0 y +. Si no hay unidad de medida, los 'saltos' tendrán diferente significado: no será lo mismo, por ejemplo, pasar de + a ++, que de ++ a +++.

Definición

La **escala de intervalo** cumple las propiedades de la escala ordinal y, además, dispone de unidad de medida.

En la escala de intervalo, las categorías se han convertido en cifras que disfrutan de una unidad que aplica por igual a todos ellos: ya tienen significado de número. Como todos 'contienen' las mismas unidades, se pueden restar entre ellos, lo que permite, por ejemplo, comparar la amplitud de varios intervalos, dando nombre a la escala.

Ejemplo 1.6: Se puede decir que entre dos cuerpos, uno a 19ºC y otro a 20ºC, hay la misma diferencia que entre uno a 29ºC y otro a 30ºC. O incluso, que el intervalo entre 10ºC y 20ºC es 5 veces mayor que entre 30ºC y 32ºC.

Ejercicio 1.3

Proponga algún otro ejemplo de variable en escala de intervalo.

1.4. Escala de razón

Historieta: *Dice "¿qué tiempo hace?" y contesta: "Ni frío, ni calor: 0º C".*

Cuando hay unidad de medida conviene preguntar si el cero es absoluto. Es decir, si el valor 0 de la escala tiene significado de "ausencia total (absoluta) de …". En la temperatura en grados centígrados, ¿significa 0º C ausencia de temperatura?

Si hay cero absoluto, entonces se está en escala de razón o de proporción.

Lectura: Kelvin relacionó la temperatura con cierta cantidad de movimiento de las partículas y encontró que éste cesaba a -273ºC, proponiendo este valor como 0 absoluto para una nueva escala de temperatura.

Ejemplo 1.7: Se puede decir que un cuerpo que está a 400º Kelvin tiene el doble de temperatura (cantidad de movimiento) que un cuerpo a 200ºK.

Ejercicio 1.4

La variable "¿tiene cefalea?" admite las categorías "nunca", "a veces" "muchas veces" y "siempre". ¿En qué escala de medida se encuentra?

Ejercicio 1.5

La variable "fracción de eyección cardíaca", ¿en qué escala está?

Nota: La escala de razón permite hacer divisiones (razones, cocientes o proporciones) entre los valores, la de intervalo también permitía divisiones pero entre las diferencias de valores, los intervalos.

En general, suele ser irrelevante distinguir entre escala de intervalo y de razón.

1.5. Escalas de medida y tipos de variables

La Tabla 1.1 resume las propiedades de las escalas de medida. Son acumulativas, ya que tener una propiedad superior requiere cumplir las anteriores. Así, una variable en escala de intervalo, además de unidad constante, tiene ordenados sus valores.

Escala	Propiedades
Nominal	Equivalencia
Ordinal	Orden
Intervalo	Unidad
Razón	Cero absoluto

Tabla 1.1 Tipos de escala y propiedades acumulativas

En el momento de escoger un tipo de análisis, se puede renunciar a propiedades superiores y utilizar uno que corresponda a las inferiores. Por ejemplo, la edad tiene unidad de medida y permite calcular la media, pero también se pueden hacer categorías (joven, adulto,…) y calcular frecuencias.

Lectura: Las escalas de medida no se deben interpretar como un proceso automático para decidir el análisis estadístico.

Otra clasificación divide a las variables en cualitativas y cuantitativas -con unidad de medida. La escala ordinal puede corresponder a ambas, ya que las propiedades de orden podrían aplicarse a categorías (como la clase social) o a expresiones numéricas (como los puntos obtenidos en una escala o 'score' como el índice de Apgar).

Otra división es en discretas o continuas. Un recuento (el número de hermanos, por ejemplo) es una variable discreta ya que sólo puede tomar un número limitado de valores. La escala nominal debe ser discreta, pero las otras escalas pueden ser tanto discretas como continuas.

Nota: No se debe confundir la naturaleza de una variable con su nivel de redondeo. Por ejemplo, aunque podemos dar la altura de forma discreta en cm, en esencia es continua.

1.6. Teoría representativa de la medida

Lectura: En este punto seguimos a Bollen. Guardia introduce el tema en la Sociedad Catalana de Estadística.

Ejemplo 1.8: ¿Podemos utilizar la edad cómo aproximación al grado de maduración? Estudiemos en qué escala de medida se encuentra. La edad que figura en el DNI estará en escala de intervalo; pero la edad como "aproximación" al grado de maduración es muy discutible: ¿representa el mismo incremento de maduración pasar de 2 a 3 años que de 42 a 43? Si la respuesta es no, al no haber unidad de medida, tampoco habrá escala de intervalo. Pero además, se podrían encontrar ejemplos de personas con menos años pero más maduras, con lo que se pondría en entredicho también la propiedad de orden. Finalmente, incluso se podría argumentar que no tienen la misma maduración dos individuos de la misma edad, con lo que ni siquiera se tendría la propiedad de equivalencia y no se podría considerar que la edad es una medida de la maduración. Pero, por otro lado, puede ser útil observar la edad de una persona para considerar qué comportamiento podemos esperar de ella. Así pues, si no se quieren perder estas posibilidades que ofrece la edad, conviene redefinir el proceso de medida.

Las escalas nominal, ordinal y de intervalo corresponden a una visión 'operativa' de la medida: se define una variable por la forma de medirla. Esta visión permitiría definir 'el cociente de inteligencia (CI)', como la variable con la que se cuantifica la inteligencia. Pero nunca permitiría definir el concepto de inteligencia —intangible en sí mismo.

Definición

Medida es el proceso que conecta un concepto con una variable latente y ésta, con variables observables.

Es decir, existe por un lado un atributo latente que no es directamente observable (por ejemplo, la inteligencia) y por otro lado, una o varias variables que pretenden cuantificar dicho atributo (por ejemplo, el CI). El CI será tanto mejor medida de la inteligencia cuanto más intensa sea su relación con la misma y menor dependencia tenga de otros factores.

La teoría representativa de la medida es estadística en el sentido de que acepta variabilidad en los resultados. Dos individuos que obtengan exactamente la misma puntuación en una prueba de

inteligencia no han de tener idéntica inteligencia, pero cabe esperar que sea más similar que la de dos casos con valores alejados.

Hay 2 propiedades que hacen a la variable observable (el CI en el ejemplo) una buena medida de la latente (la inteligencia en sí misma): son la validez y la fiabilidad. Si un proceso de medida es válida y fiable, la variabilidad de la variable observada depende exclusivamente de la variabilidad de la variable latente (el objeto de medida o el concepto latente). Al no depender de otras variables, no tendrá error sistemático y se dirá que es válida; y al no tener error aleatorio de medida, se dirá que es muy fiable o repetible.

Definición

Se dice que una variable mide de manera **válida** un concepto representado por una variable latente si está relacionada con esta variable latente y sólo con ella.

Definición

Se dice que una variable mide de manera **fiable** si sus variaciones están muy relacionadas con variaciones en el concepto —y, por tanto, dependen poco del proceso de medida.

Validez requiere ausencia de error sistemático; y fiabilidad, error aleatorio pequeño. Así, validez implica que se esté valorando el concepto y nada más: que variaciones en el concepto comporten variaciones en la medida. Por su parte, fiabilidad requiere obtener valores próximos en medidas repetidas en el mismo individuo en las mismas condiciones.

> **Ejemplo 1.9:** Los logros sanitarios en la cantidad de vida han desplazado el objetivo hacia la calidad de vida. Para muchos pacientes, es un objetivo pertinente y relevante, es decir: válido. Pero para un clínico es incómoda, ya que cambios en un mismo paciente no son explicables por variaciones en sus parámetros clínicos. Dicho de otra manera, no es fiable porque determinaciones repetidas en un paciente estable no dan la misma puntuación.

Lectura: Para saber algo más, consulte Wikipedia y las revisiones formales de Hand. En 1996 y 2002. 2002, 165: 233-261).

Ejercicio 1.6

La variable "recuento de limfocitos CD4" suele emplearse en el seguimiento del SIDA ¿En qué escala de medida se encuentra? ¿Cree que encaja en una sola escala?

Como indicador de la evolución, ¿qué opina de su validez y de su fiabilidad?

Ejercicio 1.7

El proceso de aprendizaje universitario, como unos estudios de Medicina, pretende que aquellos que lo finalicen sean capaces de ejercer como profesionales. ¿En relación a la validez y fiabilidad, qué le parece el examen MIR comparado con, por ejemplo, la observación de su trabajo delante de un paciente real?

2. Descriptiva

La escala de medida ayuda a escoger el estadístico y el gráfico para resumir los datos.

En este punto se introducirán los comandos de R que permitirán realizar un análisis descriptivo, Para ello, se empleará el conjunto de datos *'births'* del paquete *'Epi'*, que contiene los pesos de 500 recién nacidos en un hospital de Londres.

Ejemplo R

```
# Instalar y cargar Epi y cargar datos births
> install.packages('Epi')
> library(Epi)
> data(births)

# Nombres de las variables
> names(births)
[1] "id"      "bweight" "lowbw"   "gestwks"
[5] "preterm" "matage"  "hyp"     "sex"
```

La instrucción *names* aplicada a nuestro conjunto de datos, permite ver los nombres de las variables. La explicación de cada una de las variables está en la ayuda: *?births*.

2.1. Gráficos: sectores, barras y diagramas de mosaico

Lectura: José Antonio González y Lluís Jover: Cuando las relaciones entre variables son complejas o el componente aleatorio enmascara los procesos en estudio, la representación gráfica deviene una herramienta imprescindible. (…) Los gráficos, bien utilizados, permiten una aproximación nueva y enriquecedora a la información disponible.

El gráfico **de sectores** consiste en un círculo segmentado en sectores de tamaño proporcional a la frecuencia de cada uno de los valores de la variable. Este gráfico es apropiado cuando la variable toma pocos valores.

Ejemplo 2. 1: Tiempo de evolución del trastorno según grupo de pacientes:

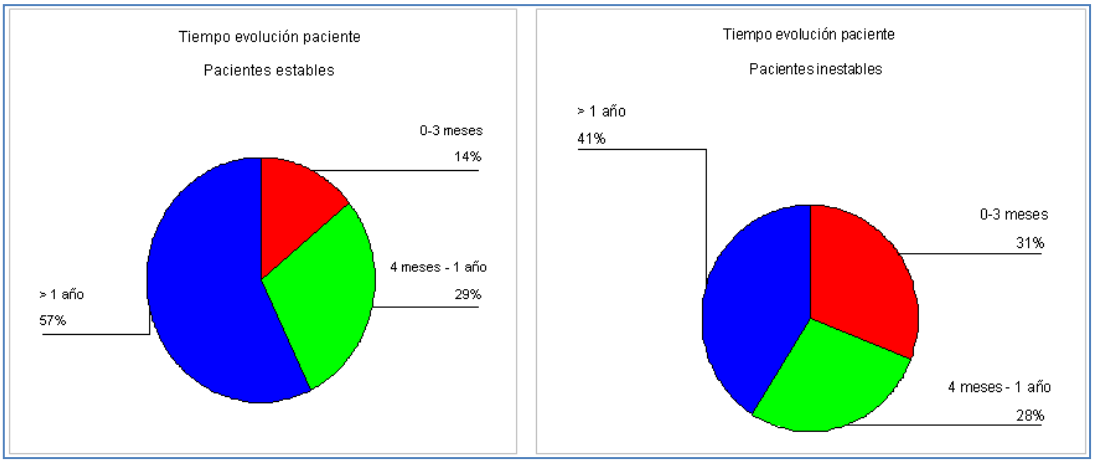


Figura 2.1 Tiempo de evolución de pacientes, estables e inestables

Historieta: Éste sí que es un buen pastel.

En R, con la instrucción *pie* puede realizar un diagrama de pastel habiendo realizado previamente la tabla de frecuencias con el comando *table*. Con los parámetros *labels* y *col* puede especificar las etiquetas y los colores del gráfico.

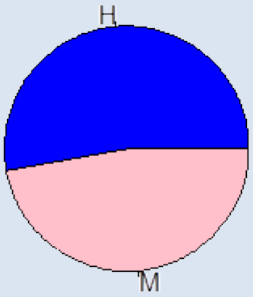
Nota: Recuerde que puede acceder a una variable de un data.frame por su nombre separado por el símbolo *$;* o bien accediendo a la posición que ocupa la columna.

Si carga los datos en memoria con la instrucción attach, podrá prescindir del nombre del conjunto de datos y bastará con escribir el nombre de la variable.

Nota: Al utilizar el *attach*, todas las variables pasan a ser objetos en memoria —lo que podría provocar ambigüedades con algún objeto con el mismo nombre. La instrucción *detach* elimina los datos de la memoria.

Ejemplo R

```
# Diagrama de sectores para la variable género
> attach (births)
> t_sex <- table(sex)
> pie(t_sex,labels=c('H','M'),col=c("blue","pink"))
```



Nota: Escribiendo *colors* ( ) en la consola se listan todos los colores disponibles en R.

De cada 12 varones, 1 no distingue el rojo del verde. Si desea aceder a la máxima población, no combine rojo y verde; mejor rojo y azul, por ejemplo.

Lectura: La ayuda de la instrucción *pie* desaconseja este tipo de gráfico: "Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data"

Emplee el diagrama **de barras** para variables discretas –nominales y ordinales. Los distintos valores se representan en el eje horizontal (abscisas) y con rectángulos de altura proporcional a la frecuencia del valor. Para que el gráfico proporcione una correcta impresión visual la escala del eje vertical (ordenadas) va desde 0 hasta, como mínimo, la frecuencia del valor modal. De no ser así, debe alertarse al lector.

Ejercicio 2.1

Rafael Ramos: "*La Figura 2.2 muestra la proporción de tratados según las distintas recomendaciones para la hipercolesterolemia.*" ¿Es un diagrama de barras?

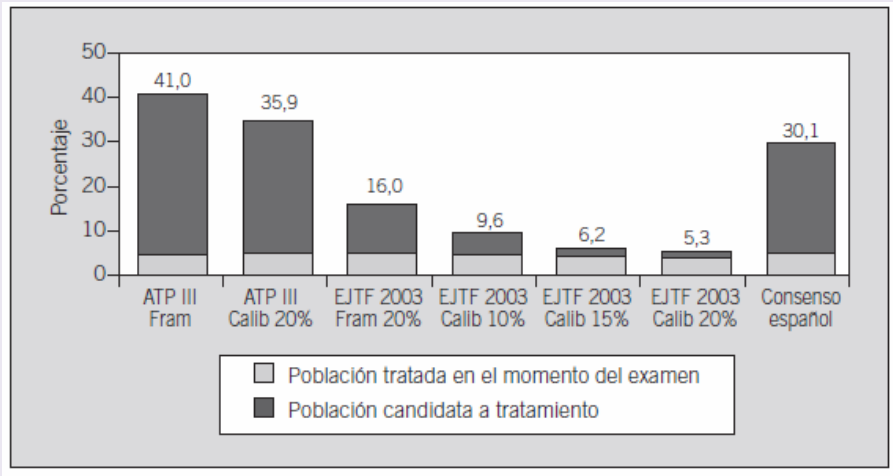


Figura 2.2 Tratamiento de la hipercolesterolemia

Una forma habitual de transmitir información errónea consiste en cambiar la escala de algún eje sin avisar al lector.

Ejemplo 2. 2: La sensación de ventaja en audiencia es diferente en los gráficos siguientes. El izquierdo no avisa del cambio de escala y engaña al lector.

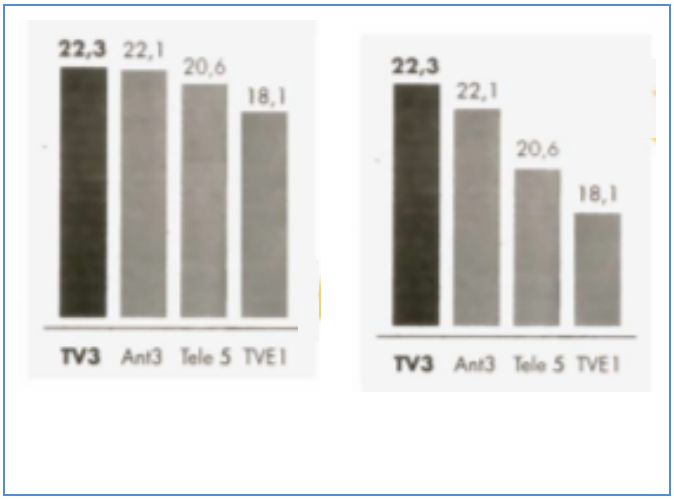


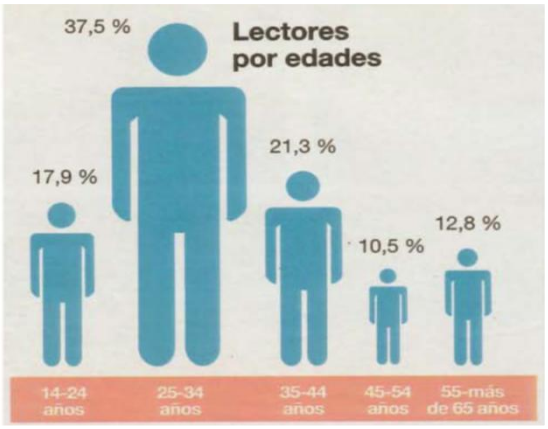
Figura 2.3 Diferente sensación por diferente escalado

Recuerde

Antes de mirar el contenido de un gráfico, lea detalladamente el pie de figura y las unidades de los ejes, observando si empiezan en 0.

Ejercicio 2.2

¿Qué opina del siguiente gráfico?



Lectura: Disfrute (en catalán) de la presentación de Pere Grima y Lluís Marco.

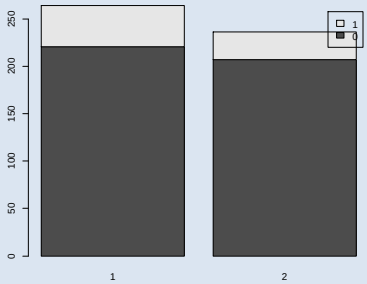
En capítulos sucesivos, con la ayuda de R, veremos gráficos más sofisticados.

Lectura: Vea la mortalidad de la campaña de Napoleón en Rusia.

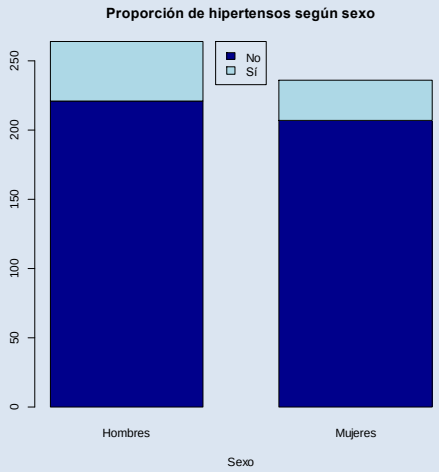
En R, el comando *barplot* realiza un diagrama de barras, siendo una tabla su primer parámetro. El argumento *legend=TRUE* añade una leyenda al gráfico. Con una tabla con 2 variables se obtiene, por defecto un gráfico de barras apiladas, pudiendose adosar las barras asignando *TRUE* al parámetro *beside*.

Ejemplo R

```
# Diagrama de barras estratificado

> T1 <- table(sex,hyp)
>colnames(T1)=c("Hombres", "Mujeres")
>rownames(T1)=c("No", "Sí")
> barplot(T1,legend=TRUE)
```



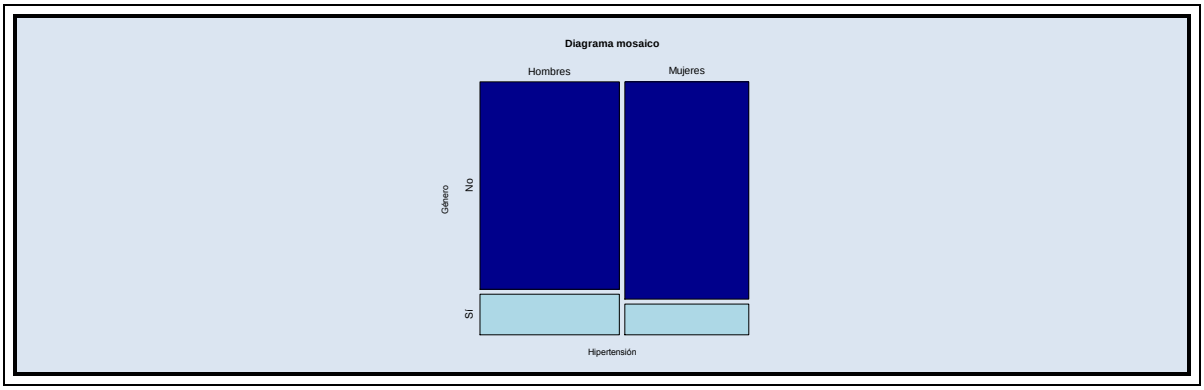
```
# Diagrama de barras estratificado mejorado

> par(las=1)
> barplot(T1,main="Proporción de hipertensos según sexo",
          col=c("darkblue","lightblue"),space=.5,
          xlab="Sexo")
#Leyenda central
>legend('top',c('No','Si'),fill=c("darkblue","lightblue"))
```



El diagrama **de mosaico** (*mosaicplot*) es parecido al diagrama de barras. La frecuencia de la primera variable (en el ejemplo, hipertensión) define la anchura de las columnas; y la de la segunda variable (género), dentro de cada categoría de la primera, define su altura. De esta forma, permite comparar las proporciones de la segunda dentro de cada categoría de la primera.

Ejemplo R

```
# Diagrama de mosaico

>rownames(T1)=c("Hombres", "Mujeres")
>colnames(T1)=c("No", "Sí")
>mosaicplot(T1,xlab="Hipertensión",ylab="Género",
           col=c("darkblue","lightblue"),main="Diagrama
           mosaico", cex.axis=1.2)
```



Nota: el carácter ~ empleado para separar las dos variables que intervienen en el *mosaicplot* se obtiene pulsando a la vez la tecla "Alt Gr" (a la derecha de 'espacio') y la tecla "4".

Ejercicio 2.3

A) Realice un *mosaicplot* de las variables peso mayor/menor de 2500 g (*lowbw*) y periodo de gestación mayor/menor a 37 semanas (*preterm*).

B) Hágalo 2 veces, intercambiando el papel de cada variable Y discuta cuál de los 2 le parece más interpretrable.

2.2. Medidas de tendencia central: moda

Las medidas de tendencia central informan dónde se sitúan las observaciones 'prototípicas'. Si las variables están en escala nominal, el parámetro más relevante para caracterizar su distribución es la frecuencia de las categorías más repetidas. En algunas ocasiones, para resumir estas variables, se las representa por su categoría más frecuente, estadístico conocido por **moda**.

Recuerde

La moda es la categoría más repetida.

Ejemplo 2.3: Miguel Martín et al para describir a los pacientes de su estudio dicen: "Los tumores de estadio II fueron los más frecuentes (55.5%)". Nótese que dan la moda pero que, además, concretan a cuántos casos representa.

En estadística, la manera de resumir toda la información contenida en una variable categórica es a través de las tablas. En R, La instrucción *table* proporciona la frecuencia de cada categoría de una variable.

Ejemplo R

```
# Frecuencia de hombres (1) y mujeres (2) en births

> table(sex)
sex
  1   2
264 236
```

Es posible hacer tablas de dos dimensiones incluyendo las dos variables categóricas separadas por una coma dentro de la instrucción *table*.

Ejemplo R

```
# Tabla de frecuencias conjuntas de Género e hipertensión

> table(sex,hyp)
hyp
sex   0   1
1   221  43
2   207  29
```

Esta tabla 2x2 contiene las frecuencias según el género del bebé (filas) y si la madre es hipertensa (0: No ; 1: Sí). La instrucción *addmargins* añade los marginales de la tabla. Primero se debe crear un objeto que contenga la tabla.

Ejemplo R

```
# Género según hipertensión materna con marginales

> T1 <- table(sex,hyp)
> addmargins(T1)
hyp
sex     0   1 Sum
1     221  43 264
2     207  29 236
Sum   428  72 500
```

La instrucción *prop.table* devuelve las proporciones de una tabla. Por defecto las calcula sobre el total; si añade un 1, sobre la fila; y si añade un 2, sobre la columna.

Ejemplo R

```
# las proporciones sobre el total deben sumar 1 todas juntas.
> prop.table(T1)             # Proporciones sobre el total
 hyp
sex     0     1
1   0.442 0.086
2   0.414 0.058
> prop.table(T1,1)     # Por fila: Cada fila suma 1.
 hyp
sex         0         1
1   0.8371212 0.1628788
2   0.8771186 0.1228814
> prop.table(T1,2)     # Por columna: Cada columna suma 1
 hyp
sex         0         1
1   0.5163551 0.5972222
2   0.4836449 0.4027778
```

2.3. Medidas basadas en el orden: cuantiles

Si las variables están en escala ordinal, es posible usar, por ejemplo, la **mediana** o valor del individuo debajo del cual se encuentra el 50% de las unidades.

Recuerde

La mediana es aquél valor que divide en dos grupos con igual frecuencia.

R cálcula la mediana con el comando *median.*

Ejemplo R

```
# Mediana del peso de los recién nacidos

> median(bweight)
[1] 3188.5
```

Las instrucciones *tapply* y *by* permiten calcular un estadístico estratificado por una variable categórica. La sintaxis es: *tapply (var. numérica, var. categórica, función).*

Ejemplo R

```
# Peso mediano de los bebés según su género

> tapply(bweight,sex,median) # 1:hombres; 2:mujeres
   1    2
3296 3107
```

Nota: Los bebés niño tienen una mediana de peso casi 200 gramos superior a los bebés niña.

Ejercicio 2.4

Obtenga la mediana de peso de los niños según si el período de gestación fue inferior o superior a 37 semanas.

Existen más medidas basadas en el orden de las observaciones. Los **cuantiles** (con 'n') son valores que dividen la población en cierto número k de grupos. El ejemplo de cuantiles más popular son los **percentiles**, que dividen la muestra en 100 partes. Los **deciles** lo hacen en 10; los **quintiles** en 5; y los **cuartiles** (con 'r') en 4.

Nótese que los cuantiles son los límites que dividen los grupos, no los grupos resultantes. Así, para dividir la muestra en cuatro partes con la misma frecuencia, bastan tres cuartiles, el 1, el 2 y el 3.

Recuerde

Hay 99 percentiles, 9 deciles, 4 quintiles y 3 cuartiles.

Ejemplo 2.4: La edad de los pacientes incluidos en un estudio tiene la distribución que muestra la figura 2.4. Por debajo de 43 años hay un 20% de las observaciones. Por tanto, el percentil 20, el 2º decil y el 1r quintil son todos ellos el mismo valor: 43 años.

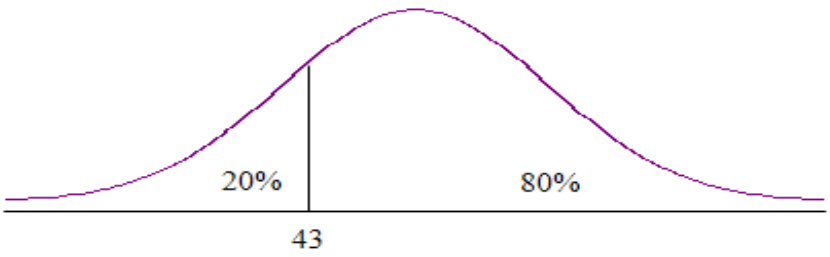


Figura 2.4. El percentil 20, el decil 2, y el quintil 1 son todos ellos 43 años

Ejercicio 2.5

La mediana, ¿a qué percentil corresponde? ¿Y a que cuartil?

Ejercicio 2.6

¿Qué percentil es el cuartil 1? ¿Y el cuartil 2? ¿Y el cuartil 3?

Los cuantiles se calculan con: *quantile ('nombre de la variable', cuantil)*

Ejemplo R

```
# Primer y tercer cuartil de los pesos
> quantile(bweight,0.25)   # 1r cuartil
25%
2862
> quantile(bweight,0.75)   # 3r cuartil
75%
3551.25
```

La instrucción *summary* proporciona un resumen de los estadísticos usuales.

Ejemplo R

```
# Descriptiva de los pesos
> summary(bweight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    628    2862    3188    3137    3551    4553
```

Nota: Las variables *gestwks* (semanas de gestación) y *preterm* (periodo de gestación inferior a 37 semanas) tienen 10 valores ausentes. Este summary no informa sobre el dato más importante: el número de casos resumido. Recuerde que R codifica los datos ausentes (*missings*) con *NA* (Notavailable).

Ejercicio 2.7

Obtenga la media, la mediana, el primer y tercer cuartil, el IQR y la desviación típica de los de los datos: 115, 117, 124, 135 y 142.

2.4. Gráficos: box-plot, histograma y otros

El **Box-plot o diagrama de caja** representa los cuartiles de variables numéricas. Los límites inferior y superior de la caja son los cuartiles 1 (percentil 25) y 3 (percentil 75). La recta interior es la mediana (cuartil 2, percentil 50). Por tanto, la caja muestra el 50% de las observaciones centrales, que podríamos considerar más "típicas". La longitud de la caja es el rango intercuartil o distancia del cuartil 1 al 3, que informa sobre el grado de dispersión. Fuera de la caja, una recta por cada lado sigue a los casos hasta llegar a la última observación, siempre que ésta tenga una distancia menor a una vez y media el rango intercuartil.

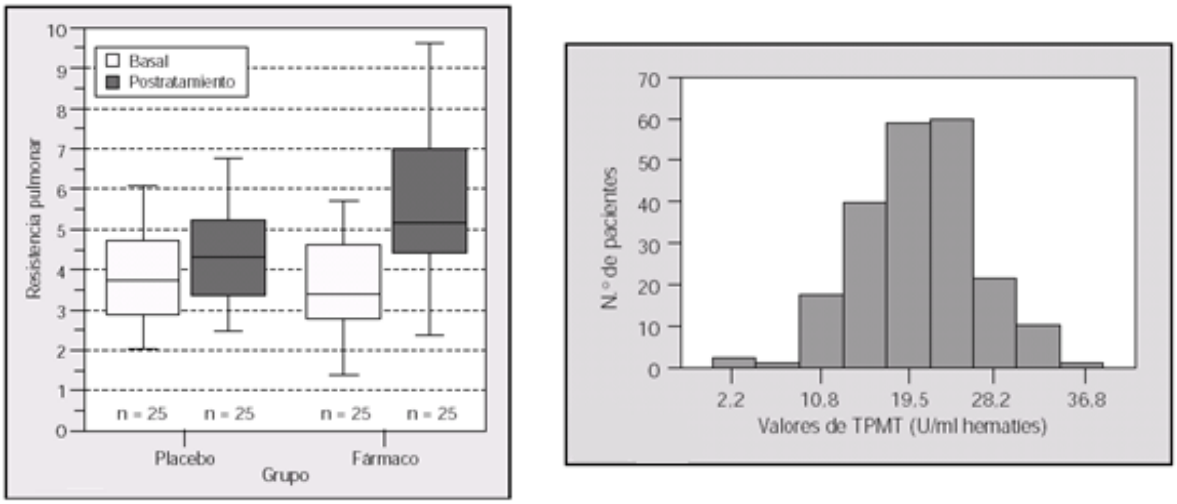


Figura 2.5. Ejemplo de box-plot y de histograma (distribución de TPMT)

Las observaciones más extremas se marcan (p.e. "*") pero no se conectan. Este gráfico es muy útil, entre otros aspectos, para valorar la simetría y detectar valores atípicos ("outliers").

Un **histograma** (Fig. 2.4 (der)) es un gráfico de variable continua dividida en intervalos de los que se eleva un rectángulo con área proporcional a su frecuencia –lo que permite intervalos de diferente amplitud.

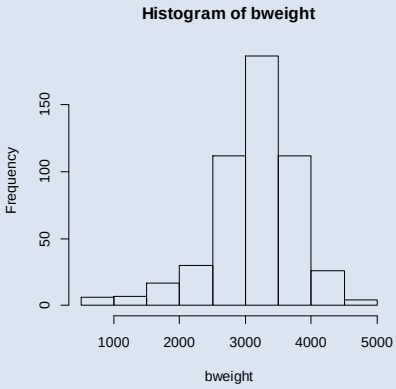
> **Nota:** Si la variable es discreta puede convenir marcarlo con rectángulos separados. Especialmente si la variable tiene muy pocos valores (p.e., número de asignaturas suspendidas").

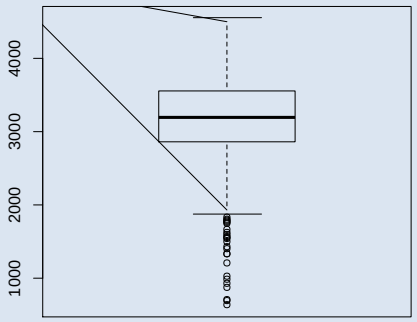
A partir de un histograma pueden construirse otros tipos de gráficos. Por ejemplo, los gráficos de línea consisten en unir con rectas los puntos medios de los intervalos contiguos, construyendo así un **polígono de frecuencias**.

Las instrucciones para realizar histogramas y diagramas de cajas son *hist* y *boxplot*, respectivamente.

Ejemplo R

```
# Gráficos de la variable peso del bebé

> hist(bweight)
> boxplot(bweight)
```

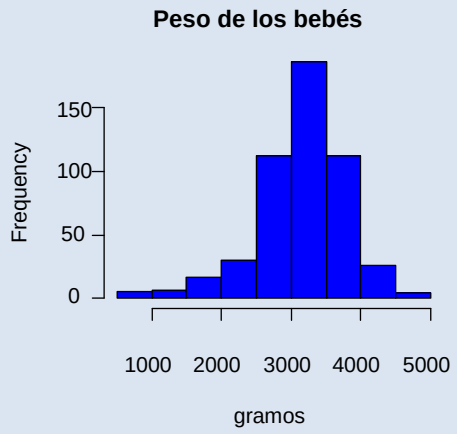


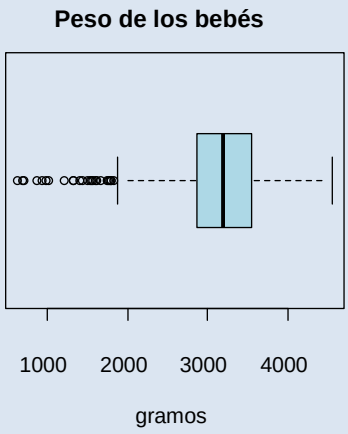


A cada uno se le pueden añadir parámetros para adaptarlos. La instrucción *par* permite fijar características a todos los gráficos.

Ejemplo R

```
# Gráficos mejorados de la variable peso del bebé

> windows(20,10)
> par (mfrow=c(1,2),las=1)
> hist(bweight,col="blue",
        main="Peso de los bebés",xlab="gramos")
> boxplot(bweight,col="lightblue",
         main="Peso de los bebés", xlab="gramos",
horizontal=TRUE)
```





Nota: La instrucción *windows (20,10)* abre una ventana de tamaño 20x10 píxeles. El parámetro *mfrow* define la posición de los gráficos en la ventana (en este caso, con 1 fila y 2 columnas); *las* indica la orientación de los números de los ejes (*las=1* los escribe siempre horizontales). Para más detalles, véase la ayuda: *?par*.

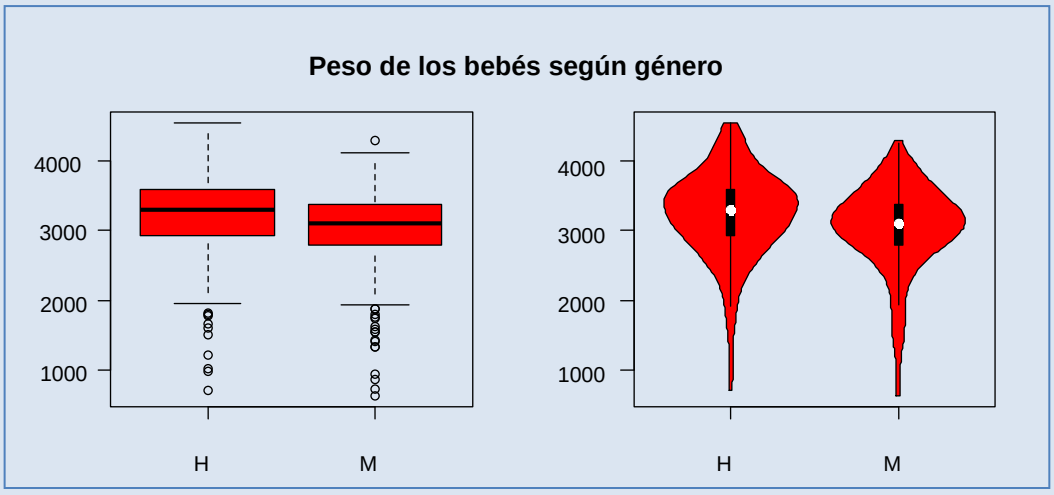
Nota: En el histograma y el boxplot, el parámetro *col* especifica el color; *main*, el título; *xlab*, la etiqueta del eje "x"; y *horizontal* dibujará el boxplot horizontal si es igual a*TRUE*. Vea más opciones con la ayuda *?hist* o *?boxplot*.

Nota: En el caso de boxplot, puede estratificar por una variable categórica añadiendo su nombre precedido de '~'.

Nota: Dispone de otros boxplots más sofisticados en otros paquetes.

Ejemplo R

```
# Peso de los bebés estratificado por el género

>install.packages('vioplot')
>library(vioplot)
>windows(20,10)
> par (mfrow=c(1,2),las=1)
> boxplot(bweight~sex,col=2,names=c("H","M"))
> vioplot(bweight[sex==1],bweight[sex==2],col=2,
                names=c("H","M"))
> title("Peso de los bebés según género",
                           outer=TRUE,line=-2)
```



Nota: Para estratificar, la sintaxis del *vioplot* es diferente, porque requiere nombres de variables diferentes para cada estrato (primero el nombre de la variable con los pesos de los bebés y luego la de las bebés). La instrucción *title* crea un título común si *outer=TRUE*. El *line = -2* coloca el título dos líneas por debajo del margen superior.

Ejercicio 2.8

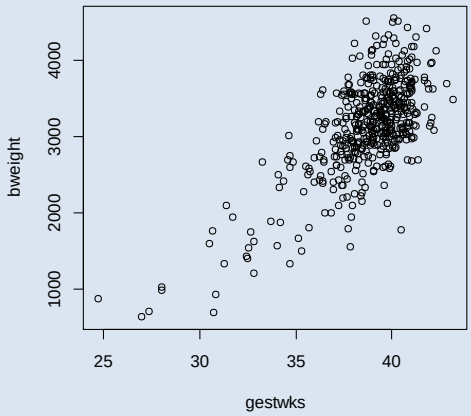
Obtenga un boxplot de las edades de las madres

2.5. Gráficos para 2 variables numéricas

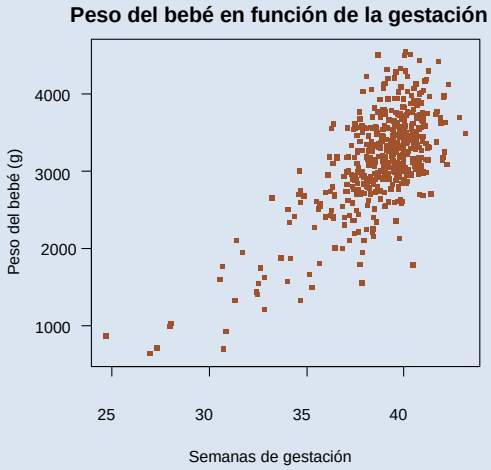
El gráfico de dispersión (*plot*) representa la relación entre dos variables numéricas.

Ejemplo R

```
# Peso del bebé según semanas de gestación

> plot(bweight~gestwks)
```




```
# Gráfico anterior (un poco más sofisticado)

> plot(bweight~gestwks,

   main="Peso del bebé en función de la gestación",
  xlab="Semanas de gestación",
  ylab="Peso del bebé(g)",
  pch=15,las=1,cex=0.6,col="sienna")
```



Con el argumento *pch* se indica el tipo de punto (el código 15 es un cuadrado sólido) y con el parámetro *cex* se indica su tamaño (por defecto, vale 1).

2.6. Medidas de tendencia central: media

Si las variables están en escala de intervalo comparten una misma unidad de medida, lo que permite sumar sus valores: es lo mismo 1 cm aportado por un individuo de 180 cm que 1cm de un señor de 150 cm. Así, para conocer el centro de la distribución puede recurrirse al promedio o media: se suman los valores obtenidos en todas las observaciones y se reparten entre el número total de casos.

> **Ejemplo 2.5**: El grupo "Asistencia Médica Integrada Continua de Cádiz" dice: "la media de pruebas por paciente es [...] menor [...] que en el grupo control".

Estudiemos la media con la ayuda de un ejemplo. Se ha preguntado a los 5 últimos pacientes que han entrado en la consulta por el número de parejas que han tenido en los últimos 48 meses y han contestado que 1, 3, 4, 5 y 7 parejas respectivamente.

La suma total de parejas es 20:

$$\begin{array}{r} 1 \\ +\ 3 \\ +\ 4 \\ +\ 5 \\ +\ 7 \\ \hline \text{suma} = \Sigma_{i=1,5}\ X_i = 20 \end{array}$$

De donde el promedio o media es de 4 parejas: $\Sigma_{i=1,5}\ X_i\ /n = 20/5 = 4$

> **Nota:** $\Sigma_{i=1,5}\ X_i$ representa la suma de los valores de la variable X en los individuos 1 a 5: es el "sumatorio desde i=1 hasta i=5 de X sub i".

En R, la media se calcula con la instrucción *mean*.

```
Ejemplos R

# Media del peso de los recién nacidos

> mean(bweight)
[1] 3136.884
```

2.7. Medidas de dispersión

Con un promedio de 4 parejas por paciente, un investigador descuidado, que se olvidara de la variabilidad, podría decir que cada uno de estos 5 pacientes ha tenido 4 parejas en los últimos 48

meses. ¡Qué sorpresa para el de 1 pareja! Y qué forma de decir mentiras. Veamos cuánto valen estas mentiras.

Dicen ellos	Se les asigna	Mentira resultante
1	4	+3
3	4	+1
4	4	0
5	4	-1
7	4	-3
Suma	20	0

Tabla 1.2 Mentira resultante si se mal-interpreta que cada paciente tiene exactamente el valor de la media

La media representa al centro de la distribución, pero ¿hasta qué punto representa a cada individuo? No todas las observaciones se sitúan en la media. Además, la diversidad puede ser riqueza. Por ello, la siguiente medida de interés estudia cuál es la distancia de las observaciones respecto la media.

Definición

La **desviación típica** o **desviación estándar** (DE) representa el alejamiento prototípico con el centro.

Hemos visto que, si se les dice que cada uno ha tenido 4 parejas, las mentiras respectivas son +3, +1, 0, -1 y -3. Ahora bien, como suman 0, el investigador descuidado podría insistir en que su cálculo es acertado, porque el promedio de sus mentiras es 0. La media, como centro de gravedad, tiene esta propiedad: se compensan los desvíos positivos con los negativos. Para evitar este efecto no deseado y poder valorar la dispersión, elevamos estas distancias al cuadrado antes de sumarlas:

Dicen ellos	Se les asigna	Mentira resultante	Mentira^2
1	4	+3	9
3	4	+1	1
4	4	0	0
5	4	-1	1
7	4	-3	9
Suma	20	0	20

Tabla 1.3 Cuadrado de la mentira si se interpreta que cada paciente tiene el valor medio

Ahora, la suma de las mentiras cuadradas es 20 parejas^2. Si las mentiras^2 que han tenido entre todos se reparten "equitativamente" en los 5 casos, se observa una "mentira^2 promedio" de 4 parejas^2, cálculo conocido por el nombre de **varianza**. Para evitar hablar de 'mentiras cuadradas' y 'parejas cuadradas' se elimina ese engorroso "cuadrado" con una raíz cuadrada, y se obtiene que la mentira

prototípica es de 2 parejas. Este valor, 2 parejas, representa la distancia o desvío (con la media) típico de todas las observaciones. Por esta razón recibe el nombre de **desviación típica**.

> **Ejemplo 2.6**: Uso de la media y de la desviación típica. Cien niños tratados han tenido fiebre durante una media de 3 días. La desviación típica (o estándar) ha sido de 1 día. Se están describiendo los resultados obtenidos en la muestra: el centro se ha situado en 3 días y los niños se alejaban de este centro, en promedio, 1 día (se entiende que se alejaban por arriba y por abajo).

Para interpretar si la desviación típica es grande o pequeña es útil el siguiente *truco*. Al ser promedio de distancias (cuadradas), habrá distancias mayores y menores, que se equilibrarán mutuamente. Así, para "compensar" a un valor que coincida exactamente con la media, es decir, que tenga un desvío igual a 0, se necesita otro valor que tenga un desvío de 2: así, *grosso modo*, los casos estarán a una distancia de 2 desviaciones típicas, tanto por encima como por debajo de la media.

> **Ejemplo 2.7**: Si la media de la fiebre era de 3 días y la desviación típica de 1 día, puede aproximarse que los niños han tenido fiebre entre 1 y 5 días.

Nota: Afinaremos este cálculo considerando la forma de la distribución.

> **Ejemplo 2.8**: Soriano et al (Med Clín 2003;121:81-5, datos redondeados): "la edad media (desviación típica) de los 11 pacientes con infección de PTC era de 70 (10) años". El centro de la distribución está en 70 años, pero no significa que todos los pacientes tengan 70 años, sino que están a su alrededor. La distancia o desviación típica que mantienen con el centro vale 10. Esta cifra representa el alejamiento "típico". En una primera aproximación, cabe imaginar que estos pacientes tienen edades comprendidas entre 50 y 90 años.

Nota: Esta aproximación puede hacerse al revés: un primer cálculo de la desviación típica en una variable simétrica, divide por 4 la distancia entre el valor más alto y el más bajo.

Recuerde

La varianza es el promedio de las distancias con la media elevadas al cuadrado. La desviación típica es su raíz cuadrada y valora el promedio de las distancias con la media: representa la distancia típica o esperada de una observación con la media.

La desviación típica muestral se representa por S. En Medicina Clínica se representa por DE (desviación estándar) y en las revistas inglesas por SD (*standard deviation*).

Ejercicio 2.9

El personal de cierto hospital camina a una velocidad media de 3km/h, siendo los extremos de velocidad 2km/h y 4km/h aproximadamente ¿Qué valor aproximado cree que puede tener la desviación típica?

Ejercicio 2.10

Los 21 pacientes con infección de la HAC tenían una edad media (DE) de 82 (8) años. Interprete la media y la desviación típica. ¿Entre qué márgenes aproximados cabe esperar que fluctúe la edad de estos pacientes?

En R, la desviación típica se obtiene con *sd* y la varianza con *var*.

Ejemplo R

```
# Medidas de dispersión de la variable bweigths

> sd(bweight)          # desviación típica (en gramos)
[1] 637.4515
> var(bweight)         # variancia (en gramos cuadrados)
[1] 406344.4
```

Nota: Note que la varianza es el cuadrado de la desviación típica.

La desviación típica es el estadístico por excelencia para valorar las dispersiones, pero requiere que exista escala de intervalo.

Nota: se ha visto que existe escala de intervalo cuando hay unidad de medida. Es decir, cuando siempre significa lo mismo un aumento de una unidad. Esta situación es verosímil cuando la variable es simétrica. Un ejemplo de asimétrica es el salario: no significa lo mismo un aumento mensual de 100€ para quien gana 500€ que para quien gana 5000€. Tampoco significa lo mismo aumentar las GOT de 10 a 40 que de 310 a 340. Las variables salario y GOT tienen una marcada asimetría, con una cola muy larga en el extremo superior (Figura 2.6). En esta situación, la desviación típica pierde sentido, ya que no puede interpretarse de la misma forma en ambas colas de la distribución.

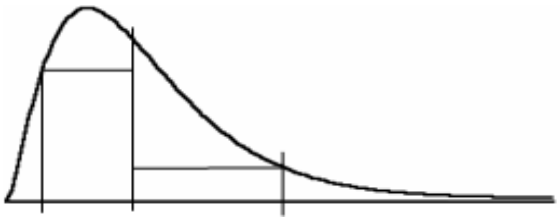


Figura 2.6. Si la distribución es asimétrica, la desviación típica no puede representar simultáneamente las distancias superiores e inferiores a la media

Recuerde

Una distribución simétrica facilita interpretar el valor de la desviación típica.

Si las variables son muy asimétricas puede renunciarse a la unidad de medida. Para valorar la dispersión en la escala ordinal hemos visto la **distancia intercuartil**.

En R, la instrucción para el cálculo de la distancia intercuartil es *IQR*.

Ejemplo R

```
# Rango intercuartílico como medida de dispersión de bweigths

> IQR(bweight)        # rango intercuartílico
 [1] 689.25
```

Escala	Propiedades	Tendencia central	Dispersión
Nominal	Equivalencia	Moda	
Ordinal	Orden	Mediana	Distancia intercuartil
Intervalo	Unidad	Media	Desv. Típica = √ Varianza

Tabla 1.4. Estadísticos apropiados según la escala de medida

La Tabla 1.4 muestra las propiedades mínimas que requiere cada estadístico. Así, por ejemplo, la media requiere escala de intervalo, pero la moda puede ser empleada en cualquier escala.

Ejercicio 2.11

Suponga que ha medido la presión arterial sistólica a 5 pacientes, 115, 117, 124, 135 y 142 mmHg.

a) Sin hacer el cálculo, diga qué valor aproximado le parece correcto para la media:

115 mmHg 125 mmHg 135 mmHg

b) Suponga ahora que el resultado observado en los 5 pacientes ha sido 100, 125, 130, 135 y 160 mmHg, con una media de 130 mmHg. Sin hacer el cálculo, diga qué valor aproximado le parece correcto para la desviación típica:

15 mmHg 20 mmHg 25 mmHg

El cálculo de la varianza presentado ha dividido por n, el número de observaciones. Pero estimar la media y la desviación típica en la misma muestra implica gastar una pieza de información, "perder un grado de libertad". El cálculo habitual de la varianza divide por "n-1" (número de casos menos uno) en lugar de por "n".

Recuerde

Divida por "n-1" al calcular la varianza.

Definición

Si x_i es el valor de la observación i-ésima y $\bar{x}$, la media muestral.

Varianza muestral $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Desviación típica muestral $S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Fórmulas abreviadas $S^2 = \frac{x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{\sum x_i^2 - \bar{x}^2 \cdot n}{n-1}$

Ejercicio 2.12 [CONSEJO: hágalo con una hoja de cálculo]

a) Suponga ahora que el resultado observado en los 5 pacientes ha sido 100, 110, 120, 130 y 140 mmHg. Calcule la media, la varianza y la desviación típica.

b) Suponga que se ha medido la presión arterial sistólica al mismo paciente 5 veces en la última visita, habiendo observado 125, 128, 130, 132 y 135 mmHg. Calcule la media, la varianza y la desviación típica.

c) ¿Cómo interpreta la diferencia de las dos S anteriores?

Digamos, para terminar, que la variabilidad no tiene porqué ser molesta. Al contrario, puede ser fuente de información o de mejora.

Ejemplo 2.9: los "errores" en la duplicación del DNA introducen ciertas variaciones que se traducen en individuos de diferentes características. La evolución de las especies se produce porque el entorno selecciona a las unidades mejor adaptadas. La selección natural precisa, por tanto, de la existencia de variabilidad.

Ejemplo 2.10: ciertas rutinas de programación generan, al azar, muchas posibles soluciones de un problema. Luego se seleccionan las mejores y se vuelve a añadir ruido para reiniciar este pequeño ciclo.

Historieta: en el paradigma de la uniformidad, las diferencias con el patrón se llaman desvíos, pero en la sociedad de la información se abre paso el paradigma biológico de la diversidad y las diferencias empiezan a ser un valor positivo. Demos pues la bienvenida a la diversidad y olvidemos las connotaciones negativas del término 'desviación'. Un término más positivo, especialmente en el ejemplo de las parejas, podría ser "diversión típica". Seguiremos buscando…

Lectura: diferente no es desviado.

2.8. Medidas de posición relativa

La existencia de diferencias representa información. El hecho de que seamos diferentes nos permite distinguirnos. Para ello, puede resultar muy útil conocer cuál es la posición de una unidad respecto a otras unidades de su entorno.

Ejemplo 2.11: Vamos a visitar a un amigo conocido en un "chat" de internet. Él vive en un poblado de África y, para identificarlo, nos ha dicho que mide 170 cm. A medida que nos acercamos a su poblado dudamos si podremos identificarlo. ¿Cuál debe ser la altura típica de su poblado? Podría ser que fueran muy altos. O todo lo contrario. Saber la media de la altura puede ser una gran ayuda. Pongamos que en su poblado dicha media sea de 150 cm. Podemos considerar "altos" a todos los que midan más de 150 y "bajos" a los que midan menos. Ahora ya sabemos que tenemos que mirar hacia los altos, pues nuestro conocido tiene una distancia positiva de 20 cm con la media del poblado.

Ahora bien, podría ser que en dicho poblado existiera una gran dispersión y nuestro conocido pasara desapercibido dentro de los altos. O podría ser que todos los habitantes estuvieran muy cerca de la media y nuestro conocido enseguida resaltara. Ahora queremos saber cuánto vale la desviación típica. Si fuera de 20 cm, nuestro conocido sería alto, pero sin destacar entre los altos: sería un "alto típico". En cambio, si la desviación típica fuera de 2 cm, sabemos que la altura de nuestro conocido resaltará mucho entre las de sus vecinos.

Definición

El procedimiento estadístico de tipificar o estandarizar el valor de una variable consiste en restarle la media y dividirlo por la desviación típica.

$$z = \textbf{desvío tipificado} = \frac{\text{valor observado - media}}{\text{desviación típica}}$$

Valores de z alrededor de 1 ó –1 representan distancias típicas al valor central. Valores cercanos a 0 representan valores muy próximos al centro de la distribución. Y valores de z mayores que 2 (o menores que –2) representan individuos que se están alejando más del doble de lo que se aleja el individuo típico.

Ejemplo 2. 11 (cont): Si la desviación típica del poblado de nuestro amigo africano es de 20 cm, el desvío tipificado de nuestro amigo vale 1:

$$z_1 = \frac{170 - 150}{20} = 1$$

En cambio, si la desviación típica del poblado fuera 2 cm, el desvío tipificado de nuestro amigo sería 10:

$$z_2 = \frac{170 - 150}{2} = 10$$

Ejercicio 2.13

En cierta población, el colesterol HDL tiene una media de 45 mg/dl y una desviación típica de 10 mg/dl. Un paciente con colesterol de 70, ¿qué desvío tipificado tiene? ¿Cómo interpreta este valor? ¿Y para un paciente con 35 mg/dl?

Ejemplo 2.11 (cont): El hipotético desvío tipificado de nuestro amigo de 1 indica que nuestro amigo es un alto típico. En cambio, el desvío de 10 indica que nuestro amigo tiene una altura atípica. Desde un punto de vista estadístico, se trata de un caso "raro", extremo.

Recuerde

Un caso que se aleje más de 2 DT está fuera de la banda ("*outlier*").

Ejemplo 2.12: Un *outlier* sería un señor que mida más de 210 cm (criterio univariante) o un señor de 180 cm que pese 55 Kg (criterio bivariante).

Nota: dónde ponemos la banda o límite es arbitrario. Evite sacar conclusiones precipitadas.

Historieta: Un caso fuera de límites (*outlier*) puede ser un elemento *extra-ordinario* que sí pertenece a esa población (Figura 2.7 (izquierda)); pero también puede ser una contaminación en la muestra (Figura 2.8 (derecha)).

Figura 2.7

Figura 2.8

Figuras 2.7 y 2.8: Dos tipos diferentes de *outlier*: el de la izquierda pertenece a la población, el de la derecha, no.

Conviene distinguir entre situaciones imposibles (p. ej., 300 cm) o situaciones raras pero posibles (p. ej., 227 cm). Un *outlier* alerta sobre posibles errores de trascripción, o posibles contaminaciones de la muestra, pero no es ninguna prueba definitiva de dato erróneo, por lo que se deben consultar y revisar estas anomalías. No se aconseja eliminar un caso por criterios de "rareza" estadística.

Ejercicio 2.14

La variable RFS tiene una media de 400 y una desviación típica de 150. Defina criterios para detectar datos "sospechosos" en las semanas 0, 6, 12 y 24 del estudio. ¿Qué hará con estos casos?.

Ejercicio 2.15

Si consulta al investigador que generó los datos, ¿cuándo le parece más oportuno?

2.9. Descripción de los participantes

El primer criterio para valorar si unos resultados aplican a nuestro entorno es mirar los criterios de elegibilidad. Pero los pacientes finalmente reclutados pueden haberse desplazado dentro de esos

criterios de selección. Por ello, hay que mirar la descripción de los pacientes incluidos, usualmente en las tablas o en el texto.

Ejemplo 2.13: Bobes: "Los 168 sujetos incluidos en el estudio (52 pacientes estables, 116 inestables) … fueron en su mayoría mujeres (85 y 82%, respectivamente), con una media (DE) de edad de 47 (12) y 45 (13) años, respectivamente, y nivel de estudios primario. En ambos grupos, la mayoría de pacientes estaba en situación laboral activa (el 35 y el 47%), si bien también fue importante el porcentaje de amas de casa incluidas (el 29 y el 35%)."

Características	**Grupo vitaminas (n = 110)**	**Grupo placebo (n=115)**
Edad media ± SD, *y*	65.3 ± 5.1	63.6 ± 4.3
Fumadores, *n (%)*	65 (59.1)	40 (34.8)
IMC media ± SD, kg/m^2	27.2 ± 3.7	25.4 ± 7.1
Tensión arterial media ± SD, *mmHg*		
Sistólica	110 ± 10	111 ± 9
Diastólica	65 ± 7	67 ± 8
Enfermedad concomitante, *n (%)*		
Hipertensión idiopática	20 (18.2)	16 (13.9)
Diabetes	13 (11.8)	7 (6.1)

Tabla 1.5 Ejemplo ficticio de tabla con características iniciales, clínicas y demográficas.

Las guías de publicación (p.e. CONSORT punto 15) explican con detalle cómo se han de presentar los datos tanto de las variables continuas como de las variables discretas.

Nota técnica: Observe que esta directriz dice que el error estándar y los intervalos de confianza (todavía no estudiados) no sirven para describir las condiciones iniciales de los casos.

Ejercicio 2.16
¿Cómo representaría los resultados de las siguientes variables?
a) Glicemia en ayuno en personas sanas
b) Transaminasas en enfermos
c) Grado de cardiopatía (nivel I a IV) según NYA
d) Presión arterial sistólica

En general, por eficiencia, las revistas sugieren dar la descriptiva detallada en tablas. Aunque permiten resaltar algo en el texto, no les gustan las repeticiones.

Variables	Pacientes estables (n = 52)	Pacientes inestables (n = 116)
Edad (años), media (DE)	47,5 (12,1)	45,2 (12,8)
Sexo		
Varones	8 (15,4)	21 (18,3)
Mujeres	44 (84,6)	94 (81,7)
Nivel de educación		
Sin estudios	3 (5,9)	8 (7,0)
Estudios primarios	33 (64,7)	72 (62,6)
Estudios secundarios	9 (17,6)	19 (16,5)
Estudios universitarios	6 (11,8)	16 (13,9)
Situación laboral		
Trabaja fuera de casa	18 (34,6)	53 (47,3)
Parado	2 (3,8)	7 (6,3)
Jubilado	2 (3,8)	3 (2,7)
Incapacidad laboral o invalidez permanente	13 (25,0)	9 (8,0)
Ama de casa	15 (28,8)	39 (34,8)
Estudiante	2 (3,8)	1
Diagnóstico (código DSM-IV)		
Trastorno depresivo mayor, episodio único (296.2)	16 (30,8)	23 (19,8)
Trastorno depresivo mayor, recidivante (296.3)	17 (32,7)	49 (42,2)
Trastorno distímico (300.4)	12 (23,1)	20 (17,2)
Trastorno adaptativo con depresión (309.0)	7 (13,5)	24 (20,7)
Tiempo de evolución del trastorno		
0-3 meses	7 (14,3)	35 (30,7)
4 meses-1 año	14 (28,6)	32 (28,1)
> 1 año	28 (57,1)	47 (41,2)
Gravedad del trastorno		
Un poco enfermo	2 (3,8)	
Levemente enfermo	22 (42,3)	7 (6,0)
Moderadamente enfermo	22 (42,3)	84 (72,4)
Gravemente enfermo	6 (11,5)	24 (20,7)
Entre los casos más graves de la enfermedad		1

Los datos se expresan como n (%) salvo cuando se indica otra cosa.

Tabla 1.6 Características sociodemográficas y clínicas de los pacientes en estudio

Soluciones a los ejercicios

1.1 La clasificación más habitual de los seres humanos es en masculino o femenino.

1.2 La clase social, en alta, media y baja es otro ejemplo de escala ordinal. Un "score", tipo test de Apgar, entre 0 y 10, cumple las propiedades de orden: un recién nacido con valor de 10 está mejor que otro con 9 y así sucesivamente.

1.3 El peso de un paciente está en escala de intervalo: la diferencia entre un sujeto con 65 y otro con 70 Kg. es la misma que la que existe entre otros dos de 85 y 90 Kg.

1.4 "¿Tiene cefalea?" está claramente en escala ordinal.

1.5 Físicamente, la fracción de eyección tiene unidad de medida y cero absoluto; pero en su interpretación clínica, como nivel de rendimiento cardíaco, no somos nosotros los que debemos decidir si significa lo mismo subir de 28 a 32%, que de 52 a 54%: un clínico debe valorar si estos cambios se interpretan de la misma forma para decidir la escala y la mejor forma de resumirla (media y SD si acepta unidad de medida, mediana y rango intercuartil, en caso contrario).

1.6 Desde el punto de vista de escala de medida, el recuento de CD4 posiblemente estaría en una escala cuantitativa de intervalo, con un mismo significado del incremento al pasar de 150 a 200 que de 550 a 600. Una primera dificultad aparece si el aparato de medida precisa un valor mínimo, pongamos 20, para poder detectar los linfocitos. Si fuera así, tendríamos que se trataría de una variable "censurada", en la que todos los valores inferiores a 20 han sido reconvertidos en un único valor "no detectado". De esta forma, se dispondría de una variable parcialmente de intervalo y parcialmente nominal u ordinal. Otra dificultad es si esta variable se pretende utilizar como indicadora de una variable subyacente, no directamente observable, como podría ser la evolución de ese paciente ante su enfermedad. ¿Aún significa lo mismo un incremento de 150 a 200 que de 550 a 600? ¿O de 375 a 425? Posiblemente no. Todo apunta a que debamos 'movernos' desde la visión operativa de la medida hacia la visión representativa y preguntarnos, no por la escala, sino por la validez y la fiabilidad. Esta última será posiblemente alta en el sentido de que, repetida la determinación de CD4 se obtienen valores similares. Pero esta fiabilidad será no tan alta si lo que se pretende que sea similar es la evolución, por lo que deberá matizarse también cómo se define la fiabilidad. En cuanto a la validez, se trata de estudiar cómo ayudan los valores de CD4 a predecir esta evolución, lo que puede estudiarse, por ejemplo, con la ayuda de términos como sensibilidad y especificidad estudiados más adelante.

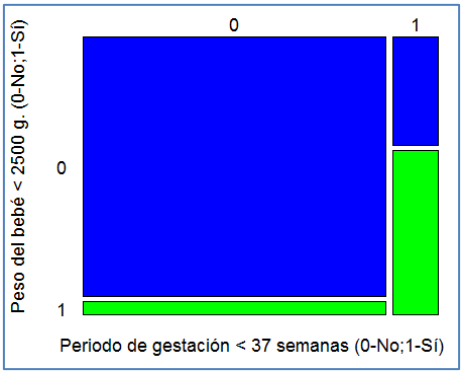
1.7 El examen MIR es menos válido, ya que mide la capacidad de contestar unas preguntas, no la de actuar profesionalmente. En cambio, es mucho más fiable, en el sentido de que si se repite la evaluación de un mismo individuo (con otras preguntas) se obtendrán puntuaciones mucho más similares (sea quien sea el evaluador) que si se cambia el paciente-caso o el examinador. [Y no olvidemos que el evaluador puede estar sometido a muchos sesgos, pero eso es quizás otra discusión.]

2.1. No. No suma el 100%. Es decir, no es el gráfico de una sola variable sino de varias: está poniendo en la misma figura el porcentaje de pacientes que cumplen cada uno de esos criterios. Como cada paciente puede tener más de uno, están recogidos en variables diferentes. En resumen, no es un histograma ni un diagrama de barras ya que éstos representan una sola variable.

2.2. Que engaña: la impresión visual del tamaño viene por el área, no por la altura. Pero en este gráfico la proporcionalidad parece ser con la altura no con el área.

2.3. A) El código para obtener el mosaicplot es

```
>par(mfrow=c(1,1),las=1, cex.lab=1.1)
>mosaicplot(preterm~lowbw, col=c("blue","green"),
            xlab="Periodo de gestación < 37 semanas (0-No;1-Sí)",
      ylab="Peso del bebé < 2500 g. (0-No;1-Sí)",
            main="",cex.axis=1.1)
```



B) Esta segunda pregunta es muy difícil. El capítulo 4 aborda a fondo está cuestión. Digamos, por ahora, que el porcentaje de bajo peso (variable posterior) según nivel de periodo (variable inicial) es más interpretable.

2.4. Se obtiene que la mediana de los bebés prematuros es más de 800 g. inferior.

```
>tapply(bweight,preterm,median)
   0    1
3282 2404
```

2.5. La mediana es el percentil 50 y el cuartil 2.

2.6. El cuartil 1 equivale al percentil 25; el cuartil 2, al percentil 50 y el cuartil 3, al percentil 75.

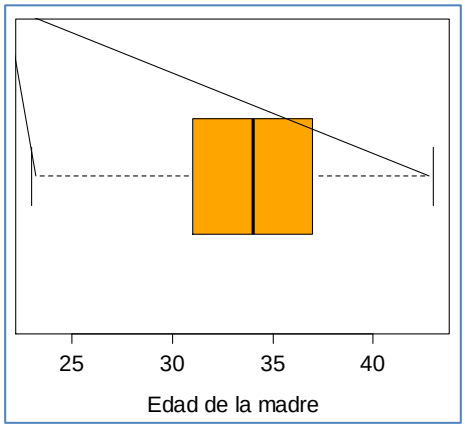
2.7. Todos los estadísticos se toman de la función summary a excepción de la desviación típica.

```
>valores<- c(115, 117, 124, 135, 142)
>summary(valores)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  115.0   117.0   124.0   126.6   135.0   142.0
>sd(valores)
[1] 11.63185
```

La media es 126.6; la mediana, 124; el primer y tercer cuartil son 117 y 135 respectivamente; el IQR vale 18 (135 – 117); y la desviación típica es 11.6.

2.8. Antes de realizar el diagrama de caja, se fijan algunos parámetros gráficos.

```
>par(mfrow=c(1,1),las=1)
>boxplot(matage, col="orange", horizontal = TRUE,
         xlab="Edad de la madre")
```



2.9. Si podemos aceptar que alguien que camina muy despacio va a 2 Km/h y alguien muy rápido a 4 Km/h, cabe esperar una desviación típica próxima al valor 0.5 Km/h, dado que (4-2)/4 es 0'5.

2.10. El doble de la desviación típica es 16, que restado y sumado de 82, da 66 y 98. Se trata de una población anciana (82 años) pero que cubre un amplio margen, ya que posiblemente fluctúa entre 66 y 98.

2.11. a) 125 (de hecho el valor exacto es 126.6 mmHg.)

b) 15 (en este caso, dados los pocos datos, la aproximación de dividir el rango entre 4 no es tan buena. El valor real es 21.5 mmHg.)

2.12. a) Media $\overline{x}$ = 120 mmHg; variancia $S^2 = \frac{(x_i - x)^2}{n-1} =$ 1000 / 4 = 250 mmHg, y desviación típica S = $\sqrt{250}$ ≈ 16 mmHg.

b) Media $\overline{x}$ = 130 mmHg; variancia S^2 = 58 / 4 ≈ 14.5 mmHg, y S = $\sqrt{14.5}$ ≈ 3.81 mmHg.

c) La desviación típica del segundo enunciado es muy inferior, ya que sólo incluye las oscilaciones debidas a las fluctuaciones intra-caso, que pueden ser debidas a cambios en el individuo pero también a errores en el procedimiento de medida. En el primer caso, además de estas oscilaciones, también aparecen las debidas .a las diferencias entre individuos.

2.13. Al paciente con un valor de 70 mg/dl le corresponde un desvío típico de +2.5, lo que indica que está por encima y de forma marcada, ya que tiene 2.5 veces la distancia habitual de los valores con la media. El paciente con un valor de 35mg/dl tiene un desvío típico de -1, lo que indica que está por debajo, pero ahora de forma típica. Estadísticamente, el primer caso podría ser considerado como un caso extremo. Ello requiere ahora una discusión clínica.

2.14. Con esta media y desviación típica, los casos deberían estar comprendidos entre:

Valores = media ± 2 desviación típica = 400 ± 2·150 ≈ 400 ± 300 = [100, 700]

Así, los valores que fueran inferiores a 100 o superiores a 700 serían 'sospechosos' de acuerdo con este criterio univariante. [De forma simple, un criterio bivariante podría establecer como sospechoso a un paciente que sufriera variaciones de su CD4 superiores al 50%.]

Estos casos deberían ser contrastados con mucho cuidado, de acuerdo con su historia clínica, a la búsqueda de posibles errores de trascripción. Si no se encuentran errores, el valor debe darse por bueno.

Al estudiar la distribución Normal veremos que este intervalo (cambiando 2 por 1.96) contiene el 95% de las observaciones si la distribución tiene forma de campana.

2.15. Por supuesto, lo más próximo al momento en el que se generó el dato. De lo contrario, puede llegar a ser imposible verificarlo.

2.16. a) Media y desviación típica, ya que por experiencia previa cabe esperar una distribución simétrica.

b) Mediana y cuartiles 1 y 3 (o percentiles 25 y 75, que son lo mismo), ya que no parece simétrica.

c) Frecuencias y porcentajes de cada nivel I-IV.

d) Media y desviación típica, ya que parece simétrica.

Y recuerde informar siempre del número n total de casos.